

# Simple response and predictor transformations to adjust for symmetric dependency in dimension reduction for visualization

LUKE A. PRENDERGAST

*Department of Mathematics and Statistics, La Trobe University*

ALEXANDRA L. GARNHAM

*Department of Mathematics and Statistics, La Trobe University*

**ABSTRACT.** In the regression setting, dimension reduction allows for complicated regression structures to be detected via visualization in a low-dimension framework. However, some popular dimension reduction methodologies fail to achieve this aim when faced with a problem often referred to as symmetric dependency. In this paper we show how vastly superior results can be achieved when carrying out response and predictor transformations for methods such as least squares and Sliced Inverse Regression. These transformations are simple to implement and utilize estimates from other dimension reduction methods that are not faced with the symmetric dependency problem. We highlight the effectiveness of our approach via simulation and an example. Furthermore, we show that ordinary least squares can effectively detect multiple dimension reduction directions. Methods robust to extreme response values are also considered.

*Key words:* cumulative slicing estimation; Ordinary Least Squares, principal Hessian directions, robust  $M$ -estimation, Sliced Inverse Regression, Sliced Average Variance Estimates

## 1 Introduction

Let  $Y \in \mathbb{R}$  denote a random univariate response and  $\mathbf{x} \in \mathbb{R}^p$  a random  $p$ -dimensional vector of predictors. Li & Duan (1989) considered the model

$$Y = f(\boldsymbol{\beta}^\top \mathbf{x}, \varepsilon) \quad (1)$$

where  $\boldsymbol{\beta}$  is an unknown  $p$ -dimensional vector of predictor coefficients,  $f$  is the unknown link function and  $\varepsilon$  is the error term that is assumed to be independent of  $\mathbf{x}$ . Of interest is the regression function  $E(Y|\mathbf{x})$  where, ideally, a plot of  $Y$  versus  $\mathbf{x}$  can reveal the form of  $f$ . However, we are limited in this sense when  $p$  is large due to our inability to visualize

objects in high dimensions. Importantly,  $Y$  depends on  $\mathbf{x}$  only through  $\boldsymbol{\beta}^\top \mathbf{x}$  so that if we could determine  $\boldsymbol{\beta}$  then we could replace the  $p$ -dimensional  $\mathbf{x}$  with the one-dimensional  $\boldsymbol{\beta}^\top \mathbf{x}$ . Our ability to explore possibilities for  $f$  would then be enhanced due to the resulting lower-dimensional framework.

In the sample setting, let  $\{y_i, \mathbf{x}_i\}_{i=1}^n$  be  $n$  sample realizations of  $Y$  and  $\mathbf{x}$  where the relationship between  $Y$  and  $\mathbf{x}$  is assumed to be of the form given in (1). Suppose that  $\boldsymbol{\beta}$  can be estimated and let this estimate be denoted  $\hat{\boldsymbol{\beta}}$ . Then the  $y_i$ 's can be plotted against the  $\hat{\boldsymbol{\beta}}^\top \mathbf{x}_i$ 's to visually determine  $f$ . Such a plot is called an *Estimated Sufficient Summary Plot* (ESSP, see, for e.g., Cook, 1998a). The focus of our work here will be to obtain good ESSP's in settings for which estimation of  $\boldsymbol{\beta}$  is difficult.

Li & Duan (1989) extended earlier works by Brillinger (1977, 1983) to show that ordinary least squares (OLS), and robust versions, can be used to estimate the direction of  $\boldsymbol{\beta}$  when the model is of the form as in (1) and under some mild conditions for  $\mathbf{x}$ . We will provide a brief review of these results in Section 2. However, for some forms of  $f$  least squares is not expected to find  $\boldsymbol{\beta}$ . Consequently we also discuss another approach, Principal Hessian Directions (Li, 1992, PHD,), which is not restricted by these particular models. In Section 3 we propose a simple transformation of the response based on an initial PHD estimate that can be used to ensure that OLS can provide a good ESSP. Simulations are provided in Section 4 which highlight that this approach can be used to obtain vastly superior estimates. Extensions are discussed in Section 5 to consider other approaches. Finally, an example is provided in Section 6 and the paper is concluded with a discussion in Section 7.

## 2 Methods

Consider the following condition commonly referred to as the *Linear Design Condition* considered by Li & Duan (1989).

**Condition 1.** For any  $\mathbf{c} \in \mathbb{R}^p$ ,  $E(\mathbf{c}^\top \mathbf{x} | \boldsymbol{\beta}^\top \mathbf{x}) = c_0 + c_1 \boldsymbol{\beta}^\top \mathbf{x}$  for some scalar constants  $c_0$  and  $c_1$ .

Li & Duan (1989) highlight that this condition is satisfied when the distribution of  $\mathbf{x}$  belongs to the family of elliptically symmetric distributions. However, there are other situations for which this holds and Hall & Li (1993) show that Condition 1 often approximately holds in practice when  $p$  is large. One also has the possibility to utilize predictor transformations to ensure that it approximately holds (see, for e.g., Fox & Weisberg, 2011).

## 2.1 Least squares and similar approaches

When Condition 1 and the model in (1) hold, Li & Duan (1989) show that the OLS slope vector, which is denoted  $\mathbf{b} = [\text{Var}(\mathbf{x})]^{-1} \text{Cov}(\mathbf{x}, Y)$ , satisfies  $c\boldsymbol{\beta}$  for a  $c \in \mathbb{R}$ . Consequently, OLS can recover the direction of  $\boldsymbol{\beta}$  when  $c \neq 0$  and a plot of  $Y$  versus  $\mathbf{b}^\top \mathbf{x}$  used to seek  $f$ . It should be pointed out that any nonzero  $c$  is adequate since any  $\mathbf{b}$  in the direction of  $\boldsymbol{\beta}$  is suitable for finding an appropriate link function. In practice, OLS can be used to obtain  $\hat{\mathbf{b}}$ , the usual OLS slope estimate, and an ESSP created using the  $y_i$ 's and the  $\hat{\mathbf{b}}^\top \mathbf{x}_i$ 's. While OLS is one simple approach, Li and Duan's results are generalized to include estimators satisfying

$$\underset{a, \mathbf{b}}{\text{argmin}} \frac{1}{n} \sum_{i=1}^n \rho(a + \mathbf{b}^\top \mathbf{x}_i, y_i) \quad (2)$$

provided that  $\rho$  is convex in its first argument and that a solution exists. Hence, other possibilities can be robust estimators such as  $M$ -estimators with the Huber weight function (Huber, 1973). While the temptation would be to only consider a robust approach when possible errors are present in the data set, Prendergast & Sheather (2013) showed that for some models the robust estimators can outperform OLS even when data is sampled without error. Similarly, Prendergast (2008) used trimming of influential observations to also improve estimates.

If an estimator is expected to find the direction of  $\boldsymbol{\beta}$ , then it is required that  $c \neq 0$ . Some discussion of when this does not occur (i.e. when  $c = 0$ ) can be found in Li (1991) and Cook & Weisberg (1991). While these discussions are for a different method they can similarly be applied to OLS. That is, when the link function  $f$  is symmetric about the mean of  $\boldsymbol{\beta}^\top \mathbf{x}$ , then  $\mathbf{b} = \mathbf{0}$ . To highlight this, we consider two simulated examples; the first does not have the symmetric dependency issue while the second does. The models we will use are

**Model 1.**  $Y = \sin(0.5\boldsymbol{\beta}^\top \mathbf{x}) + 0.05\varepsilon$

**Model 2.**  $Y = \cos(0.5\boldsymbol{\beta}^\top \mathbf{x}) + 0.05\varepsilon$

where, for both models,  $\mathbf{x} \sim N_{10}(\mathbf{0}, \mathbf{I}_{10})$ ,  $\varepsilon \sim N(0, 1)$  and  $\boldsymbol{\beta} = [1, -2, 0, \dots, 0]^\top$ .

In Figure 1 we provide true views (where the  $y_i$ 's are plotted against the ideally dimension reduced  $\mathbf{x}_i$ 's - i.e. the  $\boldsymbol{\beta}^\top \mathbf{x}_i$ 's) and ESSP's where OLS has been used as the estimator. Plots A and B are for Model 1 and Plots C and D for Model 2 where, in both cases,  $n = 100$  observations have been randomly generated. If OLS has performed well, then we would expect the ESSP to look similar to the true views with possible differences in scale on the horizontal axis since OLS is targeting  $c\boldsymbol{\beta}$  for a  $c$  that is not necessarily one. For Model 1, we can see that OLS has performed exceptionally well producing an excellent ESSP as seen in

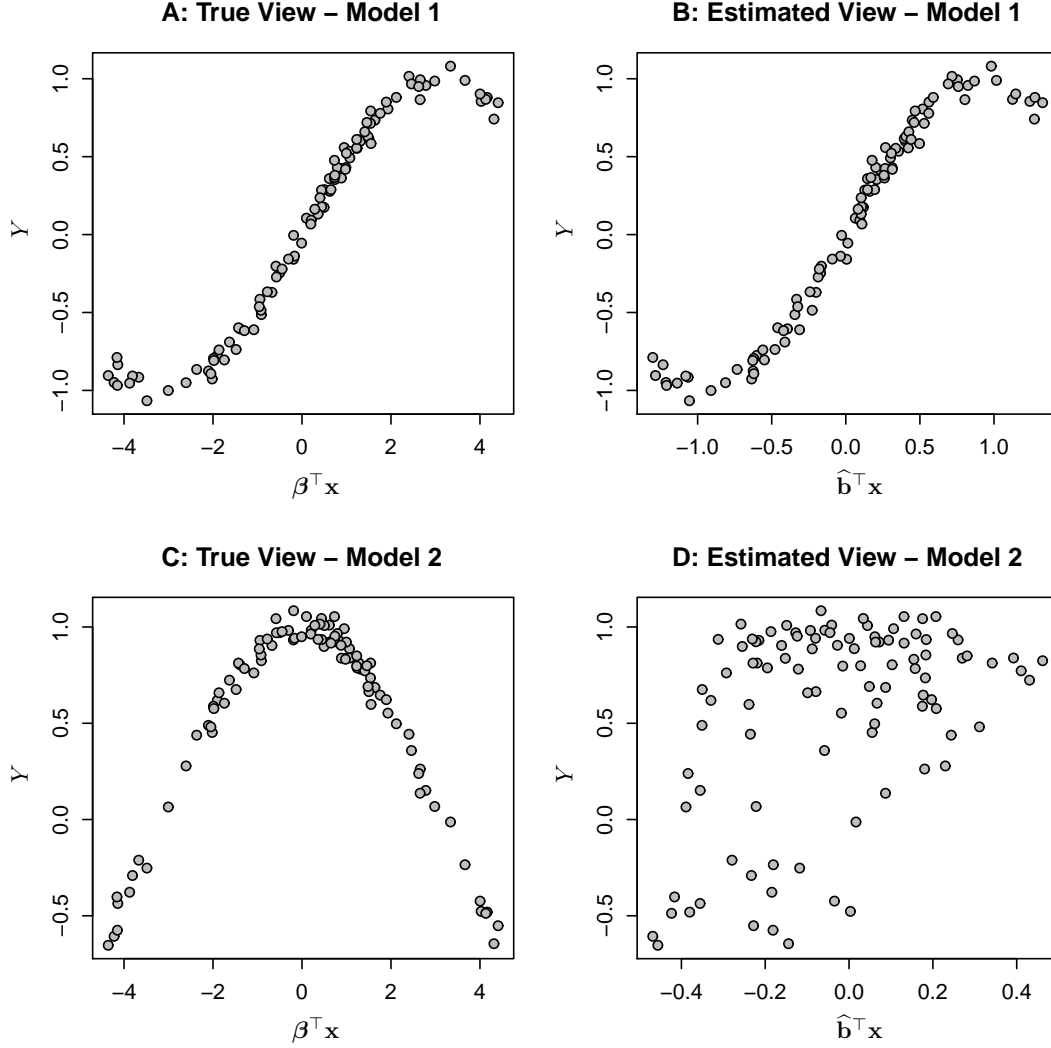


Figure 1: Plots of  $y_i$ 's versus the  $\beta^\top \mathbf{x}_i$ 's (True Views) and  $y_i$ 's versus the  $\hat{\mathbf{b}}^\top \mathbf{x}_i$ 's (Estimated Views - ESSPs) for 100 observations generated for Model 1 (Plots A and B) and Model 2 (Plots C and D). OLS was used to estimate the direction of  $\beta$ .

Plot B. However, OLS has failed for Model 2 with an ESSP in Plot D that does not provide any evidence of a relationship between the responses and dimension reduced predictors. The true view though shows that there is certainly something to find. Recall that Model 2 exhibits symmetric dependency and OLS is trying to estimate  $0 \times \beta$ .

## 2.2 Principal Hessian Directions

Li (1992) introduced Principal Hessian Directions (PHD) - a method that does not suffer

from the symmetric dependency problem and one that is also capable of finding multiple vectors of predictor coefficients. That is, the model can be assumed to be of the form

$$Y = f(\beta_1^\top \mathbf{x}, \dots, \beta_K^\top \mathbf{x}, \varepsilon) \quad (3)$$

in which case it is desirable to find a basis for the span of the  $\beta_k$ 's.

Consider the following condition required by PHD.

**Condition 2.**  $\mathbf{x} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ .

When Condition 2 holds imposing normality of the predictor, Condition 1 also holds. As a consequence, if PHD is applicable due to this condition being met then so to is OLS. There are slightly weaker conditions for PHD to work, namely that  $\text{Var}(\mathbf{x}|\beta^\top \mathbf{x})$  is constant in conjunction with assuming Condition 1 holds. Recent work by Leeb (2013) show that this will often hold approximately in practice.

While OLS returns an estimate of  $\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}_{xy}$  (where  $\boldsymbol{\Sigma}_{xy}$  is the covariance vector between  $\mathbf{x}$  and  $Y$ ) as an estimate of the direction of  $\beta$ , PHD instead carries out an eigen-decomposition of an estimate to  $\bar{\mathbf{H}} = \boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}_{yxx}\boldsymbol{\Sigma}^{-1}$  where

$$\boldsymbol{\Sigma}_{yxx} = E \left[ \{Y - E(Y)\}(\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^\top \right].$$

For *many models* satisfying (1) and when Condition 2 holds, the rank of  $\bar{\mathbf{H}}$  is one and the eigenvector corresponding to the non-zero eigen-value is in the same direction as  $\beta$ . We have emphasized many models here since PHD will not be able to find the direction of  $\beta$  when there is odd symmetric dependency between  $Y$  and the mean of  $\beta^\top \mathbf{x}$ . Here odd symmetric dependency refers to the type of symmetry seen for Model 1 (see Figure 1). Li (1991) also noted that  $Y$  can be replaced by the OLS residual without changing  $\bar{\mathbf{H}}$  where, notationally, we replace  $\boldsymbol{\Sigma}_{yxx}$  by  $\boldsymbol{\Sigma}_{rxx}$  to distinguish between the two approaches. While  $\bar{\mathbf{H}}$  does not change, the estimator is influenced. Empirical and theoretical results have suggested that this residual based PHD approach is often a better estimator of  $\beta$  (Cook, 1998b; Prendergast & Smith, 2010). Consequently the residuals-based PHD will be our method of choice.

Since  $\boldsymbol{\Sigma}_{rxx}$  is moment-based, estimation is straightforward. Let  $r_1, \dots, r_n$  denote the usual OLS residuals for the regression of the  $y_i$ 's on the  $\mathbf{x}_i$ 's and also let  $\bar{\mathbf{x}}$  be the sample mean of the  $\mathbf{x}_i$ 's. Then the estimate to  $\boldsymbol{\Sigma}_{rxx}$  is

$$\hat{\boldsymbol{\Sigma}}_{rxx} = \frac{1}{n} \sum_{i=1}^n r_i (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^\top.$$

Similarly to OLS estimation, Lue (2001) has previously shown that trimming can improve PHD estimation.

### 3 Predictor and response transformations to remove symmetric dependency

Garnham & Prendergast (2013a,b) show that response transformations can greatly improve OLS and PHD estimates. Their results, however, do not solve the issue of the symmetric dependency that troubles OLS. The aim here is to introduce two transformation functions, one for the response and the other for the predictor, that can be useful in the symmetric dependency setting.

#### 3.1 Theory

The response transformation that we will be focusing on is

$$t_y(Y; \mathbf{v}) = \begin{cases} Y, & \mathbf{v}^\top \mathbf{x} > \mathbf{v}^\top \boldsymbol{\mu} \\ Y - 2[Y - E(Y|\mathbf{v}^\top \mathbf{x} = \mathbf{v}^\top \boldsymbol{\mu})], & \mathbf{v}^\top \mathbf{x} \leq \mathbf{v}^\top \boldsymbol{\mu} \end{cases} \quad (4)$$

where  $\mathbf{v}$  needs to be chosen. If  $\mathbf{v} = c_1 \boldsymbol{\beta}$  for a nonzero scalar  $c_1$ , then  $t_y(Y; \mathbf{v})$  and  $\mathbf{x}$  still satisfy the model in (1) and Condition 1. An estimator of  $\boldsymbol{\beta}$  is then the OLS slope vector estimator for the regression of  $t_y(Y; \mathbf{v})$  on  $\mathbf{x}$ . Soon we will show that in the empirical setting, good estimates to  $\boldsymbol{\beta}$  used in the transformation function can generate improved results. We also consider the following predictor transformation function

$$t_x(\mathbf{x} - \boldsymbol{\mu}; \mathbf{v}) = \text{sign}(\mathbf{v}^\top \mathbf{x} - \mathbf{v}^\top \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu}). \quad (5)$$

In Figure 2 we show the effects of the transformations on the curves defining  $Y$  in Model 2 under the assumption of zero error. In Plot A it is clear now that the transformed  $Y$  is no longer symmetric about the mean of  $\boldsymbol{\beta}^\top \mathbf{x}$  (zero). In Plot B it is clear that  $Y$  is not symmetric about the mean of  $\boldsymbol{\beta}^\top t_x(\mathbf{x} - \boldsymbol{\mu}; \mathbf{v})$ , alleviating the symmetric dependency problem. For this latter plot, the transformation folds the curve back on to itself (left-to-right).

In the theorem below, we identify that for certain choices of  $\mathbf{v}$ , the transformation in (5) can be used to find the direction of  $\boldsymbol{\beta}$ . The proof is in the Appendix.

**Theorem 1.** *Consider the predictor transformation considered in (5) and let  $\mathbf{v} = c_1 \boldsymbol{\beta}$  for any  $c_1 \in \mathbb{R}$ . Under the model in (1) and Condition 1,*

$$[Var(\mathbf{x})]^{-1} Cov[t_x(\mathbf{x} - \boldsymbol{\mu}; \mathbf{v}), Y] = c_2 \boldsymbol{\beta} \quad (6)$$

$$\{Var[t_x(\mathbf{x} - \boldsymbol{\mu}; \mathbf{v})]\}^{-1} Cov[t_x(\mathbf{x} - \boldsymbol{\mu}; \mathbf{v}), Y] = c_3 \boldsymbol{\beta} \quad (7)$$

for constant scalars  $c_2, c_3 \in \mathbb{R}$ .

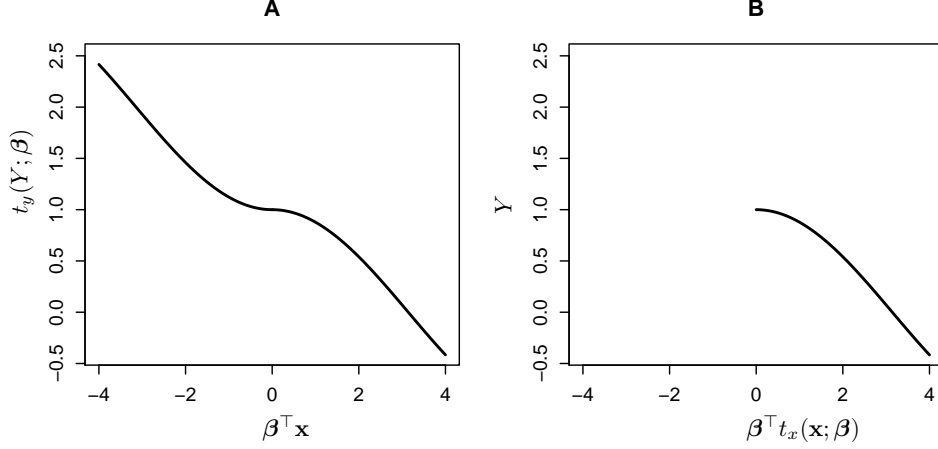


Figure 2: Under the assumption of zero error in the Model 2 and choosing  $\mathbf{v} = \beta$ , Plot A provides the plot of the transformed  $Y$  versus  $\beta^T \mathbf{x}$  and Plot B provides the plot of  $Y$  versus the transformed  $\beta^T t_x(\mathbf{x} - \mu; \mathbf{v})$ .

The second estimator provided in (7) is simply the OLS slope from the regression of  $Y$  on  $t_x(\mathbf{x} - \mu; \mathbf{v})$ . The first, provided in (6), is similar although utilizes the variance estimator for the original  $\mathbf{x}$ .

### 3.2 Application in practice

Recall that our sample of  $n$  observations are denoted  $\{y_i, \mathbf{x}_i\}_{i=1}^n$ . Throughout let  $\bar{y}$ ,  $\bar{\mathbf{x}}$ ,  $\mathbf{S}_x$  and  $\mathbf{S}_{xy}$  denote the sample mean of the  $y_i$ 's, sample mean of the  $\mathbf{x}_i$ 's, sample covariance matrix of the  $\mathbf{x}_i$ 's and the sample covariance between the  $\mathbf{x}_i$ 's and  $y_i$ 's respectively. Also let  $\mathbf{X}$  denote the  $n \times p$  design matrix whose  $i$ th row is  $\mathbf{x}_i$ .

There are two points that need clarification prior to application in practice. Firstly, how to choose an appropriate vector  $\mathbf{v}$ ? Our simulations indicate that for many models, OLS is a better estimator of  $\beta$  in (1) than PHD. However, PHD is preferred when symmetric dependency is evident in which case OLS can struggle to find  $\beta$ . Consequently, in practice we propose to set  $\mathbf{v} = \hat{\mathbf{b}}_{phd}$  - the PHD estimate to  $\beta$ . In the next section our results show that reasonable PHD estimates of  $\beta$  can lead to much-improved estimates to  $\beta$  when OLS is employed following the transformations in Section 3.1.

Secondly, the response transformation requires estimation of  $E(Y|\mathbf{v}^T \mathbf{x} = \mathbf{v}^T \mu)$ . We propose to find an approximation to this estimate as

$$\bar{y}(\mathbf{v}) = \frac{1}{m} \sum_{j \in I_m} y_j \quad (8)$$

where  $I_m$  is the set of indices for the closest  $m$   $\mathbf{v}^\top \mathbf{x}_i$ 's to  $\mathbf{v}^\top \bar{\mathbf{x}}$ . In what follows we arbitrarily set  $m = 10$  and obtain good results.

The transformations we will employ are then

$$y_i^* = t_{y_i}(y_i; \hat{\mathbf{b}}_{phd}) = \begin{cases} y_i, & \hat{\mathbf{b}}_{phd}^\top \mathbf{x}_i > \hat{\mathbf{b}}_{phd}^\top \bar{\mathbf{x}} \\ y_i - 2[y_i - \bar{y}(\hat{\mathbf{b}}_{phd})], & \hat{\mathbf{b}}_{phd}^\top \mathbf{x}_i \leq \hat{\mathbf{b}}_{phd}^\top \bar{\mathbf{x}} \end{cases} \quad (9)$$

and

$$\mathbf{x}_i^* = t_{x_i}(\mathbf{x}_i - \bar{\mathbf{x}}; \hat{\mathbf{b}}_{phd}) = \text{sign}(\hat{\mathbf{b}}_{phd}^\top \mathbf{x}_i - \hat{\mathbf{b}}_{phd}^\top \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}}) \quad (10)$$

where we will use the notations  $y_i^*$  and  $\mathbf{x}_i^*$  for convenience. The methods we will use are:

**Method 1.** *The OLS slope vector for the regression of the  $y_i^*$ 's on the  $\mathbf{x}_i$ 's.*

**Method 2.** *The OLS slope vector for the regression of the  $y_i$ 's on the  $\mathbf{x}_i^*$ 's.*

Potentially, a combination of the transformations could also be used. However, our simulations revealed the better results are achieved by using only one at a time. Additionally, another possibility exists and that is to use  $\mathbf{S}_x^{-1} \mathbf{S}_{xy}^*$ . However, our simulations also revealed that this approach was very typically inferior to the other two. For brevity, we therefore do not consider this approach further.

### 3.3 An iterative approach

A potential problem with the transformation methods is that the initial estimated direction is poor. However, one approach is to alleviate this is to apply an iterative scheme which starts with the initial estimate, obtains a new estimate after transformation and iteratively uses the new estimate as the initial estimate until convergence. Hence, a general algorithm for this approach is:

**Step 0.1:** Estimate the direction of  $\beta$  using PHD and denote this as  $\hat{\mathbf{b}}^{(1)}$ .

**Step 0.2:** Set  $i = 1$  and `tol.met = FALSE`.

**Step  $i$ :** While `tol.met` is `FALSE` do

**Step  $i.1$ :** Apply **Method  $j$**  using  $\hat{\mathbf{b}}^{(i)}$  as the direction for transformation and obtain a new estimated direction  $\hat{\mathbf{b}}^{(i+1)}$ .

**Step  $i.2$ :** If  $1 - \text{cor}^2(\mathbf{X}\hat{\mathbf{b}}^{(i)}, \mathbf{X}\hat{\mathbf{b}}^{(i+1)}) < \text{tol}$  then set `tol.met` to `TRUE`.

**Step  $i.3$ :** Increment  $i = i + 1$ .



**Step  $i + 1$ :** Return  $\hat{\mathbf{b}}^{(i)}$  as the final estimate to the direction of  $\boldsymbol{\beta}$ .

We have chosen  $\text{cor}^2(\mathbf{X}\hat{\mathbf{b}}^{(i)}, \mathbf{X}\hat{\mathbf{b}}^{(i+1)})$  as the criterion for exiting the iterating loop since, when there is correlation present amongst the columns of  $\mathbf{X}$ , notably different directions can result in very similar ESSP's - which is the targeted estimate. However, substantial differences in the squared correlation will be similarly be noted by changes in the ESSP. Such an assessment is often used; for example, Li (1991) uses the squared trace correlation which is a multi-index version to compare collections of estimated directions in dimension reduction.

## 4 Simulations

In this section we consider the performance of the transformation approaches referred to as Methods 1 and 2 defined earlier. Comparisons are also made with standard OLS and PHD estimation before we consider other methods in the next section.

Method	$p = 10$				$p = 20$			
	$n = 50$	$n = 100$	$n = 200$	$n = 500$	$n = 50$	$n = 100$	$n = 200$	$n = 500$
OLS	0.220	0.213	0.211	0.210	0.130	0.123	0.122	0.122
	<i>0.219</i>	<i>0.211</i>	<i>0.208</i>	<i>0.207</i>	<i>0.150</i>	<i>0.140</i>	<i>0.137</i>	<i>0.137</i>
PHD	0.815	0.921	0.964	0.987	0.456	0.792	0.915	0.970
	<i>0.103</i>	<i>0.041</i>	<i>0.018</i>	<i>0.006</i>	<i>0.197</i>	<i>0.082</i>	<i>0.030</i>	<i>0.010</i>
M1	0.948	0.990	0.997	0.999	0.609	0.950	0.991	0.997
	<i>0.078</i>	<i>0.008</i>	<i>0.002</i>	<i>0.001</i>	<i>0.273</i>	<i>0.050</i>	<i>0.005</i>	<i>0.001</i>
M1-it	0.975	0.994	0.997	0.999	0.683	0.985	0.994	0.998
	<i>0.062</i>	<i>0.003</i>	<i>0.001</i>	<i>0.001</i>	<i>0.293</i>	<i>0.028</i>	<i>0.002</i>	<i>0.001</i>
M2	0.948	0.989	0.996	0.999	0.613	0.948	0.989	0.997
	<i>0.076</i>	<i>0.010</i>	<i>0.002</i>	<i>0.001</i>	<i>0.271</i>	<i>0.052</i>	<i>0.006</i>	<i>0.001</i>
M2-it	0.976	0.997	0.999	1.000	0.643	0.988	0.997	0.999
	<i>0.069</i>	<i>0.002</i>	<i>0.001</i>	<i>0.000</i>	<i>0.283</i>	<i>0.034</i>	<i>0.001</i>	<i>0.000</i>

Table 1: Average  $\text{cor}^2(\mathbf{X}\boldsymbol{\beta}, \mathbf{X}\hat{\mathbf{b}})$  across 10,000 simulated runs for Model 2 with different choices of  $n$  and  $p$  and where  $\hat{\mathbf{b}}$  is the estimate from one of five methods; OLS, PHD, Method 1 (M1) and Method 2 (M2). M1-it and M2-t refer to the iterative estimation scheme for Methods 1 and 2. Standard deviations are in italics.

In Table 1 we provide simulated average squared correlations (with standard deviations in italics) for Model 2 over 10,000 runs between  $\mathbf{X}\boldsymbol{\beta}$  and  $\mathbf{X}\hat{\mathbf{b}}$  - the true and estimated dimension reduced predictors. The estimators considered are OLS, PHD and the transformation

approaches. Both  $p = 10$  and  $p = 20$  were considered. As expected OLS performs poorly due to the symmetric dependency evident in the model while PHD performs well. However, Methods 1 and 2 perform exceptionally well having successfully drawn on the good PHD estimates to remove the symmetric dependency. The iterative estimation methods also provide improved estimates. We exited the iterative procedure when  $\text{cor}^2(\mathbf{X}\hat{\mathbf{b}}^{(i)}, \mathbf{X}\hat{\mathbf{b}}^{(i+1)}) \geq 0.999$  or when ten iterations were reached. The small standard deviations for the methods indicate consistently excellent results.

Prendergast & Sheather (2013) found that robust least squares regression methods can provide improved single-index model estimates even when the data is well-behaved in the sense it has been sampled from a single index model and with a normal  $\mathbf{x}$ . Similarly, Prendergast (2008) found that trimming observations in the estimation step can also improve outcomes. We further explore this by considering the following model:

**Model 3.**  $Y = 1/|\boldsymbol{\beta}^\top \mathbf{x}| + 0.05\varepsilon$  with  $p = 20$  and  $\boldsymbol{\beta} = [1, -2, 0, \dots, 0]^\top$ .

For this model we will assume that  $\mathbf{x} \sim N(\mathbf{0}, \mathbf{I}_{20})$  so that this model also consists of the symmetric dependency that troubles OLS. Also, data simulated from the model can result in exceptionally extreme responses since the denominator in the first term on the right hand side can be very close to zero.

In Table 2 we report the average squared correlations between the true and estimated dimension reduced predictors for 10,000 simulated runs from Model 3 with standard deviations in italics. For PHD, the very large response values often generated can result in extremely poor results. However, for OLS Garnham & Prendergast (2013a) showed that using the rank of the response instead of the response itself could provide improved results. Consequently, we used the rank of the response values for PHD and this approach provides vast improvements. Therefore, the PHD results presented in this table are based on this estimation. As well as employing OLS, PHD based on ranks and Methods 1 and 2, we also consider other variations that can be used to limit the influence of very large response values. RR, RM1 and RM2 refer to the usual OLS, Methods 1 and 2 but where OLS has been replaced with the  $M$ -estimation robust version (Huber, 1964, 1973) with the Huber weight function. To do this we used the `r1m` function from the **MASS** package (Venables & Ripley, 2002) in R (R Core Team, 2013). M1-trim and M2-trim refer to Methods 1 and 2 where 10% of observations with the largest Cook's distance have been trimmed prior to the least squares step (this is one of the trimming procedures from Prendergast, 2008). The iterative estimation scheme for this model and methods did not provide improved results so for simplicity they have not been considered here. Not surprisingly, OLS and the  $M$ -estimation equivalent completely fail even for large  $n$  due to symmetric dependency. Methods 1 and 2 perform much better but

Method	$n = 100$	$n = 200$	$n = 500$	Method	$n = 100$	$n = 200$	$n = 500$
OLS	0.006	0.004	0.002	M3	0.202	0.229	0.259
	<i>0.012</i>	<i>0.009</i>	<i>0.005</i>		<i>0.150</i>	<i>0.147</i>	<i>0.150</i>
RR	0.034	0.035	0.034	RM1	0.867	0.970	0.989
	<i>0.049</i>	<i>0.049</i>	<i>0.048</i>		<i>0.130</i>	<i>0.015</i>	<i>0.004</i>
PHD	0.801	0.947	0.985	RM2	0.875	0.961	0.981
	<i>0.131</i>	<i>0.022</i>	<i>0.005</i>		<i>0.126</i>	<i>0.017</i>	<i>0.006</i>
M1	0.448	0.638	0.762	M1-trim	0.771	0.933	0.970
	<i>0.219</i>	<i>0.328</i>	<i>0.294</i>		<i>0.201</i>	<i>0.082</i>	<i>0.031</i>
M2	0.448	0.594	0.661	M2-trim	0.722	0.874	0.922
	<i>0.1219</i>	<i>0.205</i>	<i>0.193</i>		<i>0.172</i>	<i>0.081</i>	<i>0.049</i>

Table 2: Average  $\text{cor}^2(\mathbf{X}\boldsymbol{\beta}, \mathbf{X}\hat{\mathbf{b}})$  across 10,000 simulated data sets for Model 3 with different choices of  $n$  and  $p$  and where  $\hat{\mathbf{b}}$  is the estimate from various methods. RR, RM1 and RM2 refer to Methods OLS, M1 and M2 but where robust regression  $M$ -estimation has been used in the regression step with the Huber weight function. M1-trim and M2-trim refer to Methods M1 and M2 but where 10% of observations with the largest Cook’s distance were trimmed. Standard deviations are in italics.

can still struggle as evident by the moderate average squared correlations and large standard deviations. On the other hand PHD performs well, in particular for the larger sample size settings. For Methods 1 and 2 coupled with  $M$ -estimation, we see improved performance over PHD for both methods. These results suggest that by using the good PHD results in the transformation step to remove the symmetric dependency problem and then  $M$ -estimation to protect against large response values, excellent results can be achieved. For the trimming approaches, improvements have been found when compared to standard Methods 1 and 2, however the results are a little worse than PHD and much worse than the transformation plus  $M$ -estimation methods.

## 5 Inverse regression methods and multiple direction OLS

### 5.1 Inverse regression approaches

The transformations discussed in Section 3.2 are certainly not limited to OLS and PHD. Here we briefly discuss the use of Sliced Inverse Regression (SIR, Li, 1991) and Sliced Average

Variance Estimates (SAVE, Cook & Weisberg, 1991). For brevity, we only briefly discuss these methods here and the reader is directed to the aforementioned articles for more detail. Let  $S_1, \dots, S_H$  denote  $H$  non-overlapping yet collectively exhaustive intervals covering the range of  $Y$ . Let  $\boldsymbol{\mu}_h = E(\mathbf{x}|Y \in S_h)$  ( $h = 1, \dots, H$ ) denote slice means and consider the matrix  $\mathbf{V} = \boldsymbol{\Sigma}^{-1/2} \sum_{h=1}^H p_h (\boldsymbol{\mu}_h - \boldsymbol{\mu})(\boldsymbol{\mu}_h - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1/2}$  where  $p_h$  is the probability of  $Y \in S_h$ . For  $\boldsymbol{\gamma}$  denoting an eigenvector of  $\mathbf{V}$  corresponding to a nonzero eigenvalue, Li (1991) showed that  $\boldsymbol{\Sigma}^{-1/2} \boldsymbol{\gamma}$  is an element of the span of  $\boldsymbol{\beta}_k$ 's from the model in (3) provided a  $K$ -direction version of Condition 1 holds. Consequently, if  $\mathbf{V}$  is rank  $K$  then SIR can recover a complete basis for the dimension reduction directions. However, SIR suffers from the same problems with symmetric dependency as OLS (Li, 1991; Cook & Weisberg, 1991) and is therefore a candidate for the same type of transformation.

Cook & Weisberg (1991) introduced SAVE which does not suffer from symmetric dependency issues. It does, however, and additional to Condition 1 require that  $\text{Var}(\mathbf{x}|\boldsymbol{\beta}_1^\top \mathbf{x}, \dots, \boldsymbol{\beta}_K^\top \mathbf{x})$  is constant. Both conditions are satisfied when  $\mathbf{x}$  is normally distributed although both will often approximately hold in practice (Hall & Li, 1993; Leeb, 2013). Let  $\boldsymbol{\Sigma}_h = \text{Var}(\mathbf{x}|Y \in S_h)$  ( $h = 1, \dots, H$ ) denote slice covariance matrices. Then SAVE is carried out similar to SIR but where  $\mathbf{M} = \sum_{h=1}^H p_h (\mathbf{I}_p - \boldsymbol{\Sigma}^{-1/2} \boldsymbol{\Sigma}_h \boldsymbol{\Sigma}^{-1/2})^2$  is used instead of  $\mathbf{V}$ . For some models SAVE requires large sample sizes to achieve good results. However, we have found that a variation of SAVE that was proposed by Zhu *et al.* (2010) called Cumulative Variance Estimation (CUVE) often provides excellent results. For  $\mathbf{z} = \boldsymbol{\Sigma}^{-1/2}(\mathbf{x} - \boldsymbol{\mu})$ , CUVE estimates  $E[\{P(Y \leq \tilde{Y})\mathbf{I}_p - \text{Var}(\mathbf{z} I(Y \leq \tilde{Y}))\}^2]$  where  $\tilde{Y}$  is an independent copy of  $Y$  and where  $I(\cdot)$  is the indicator function taking the value 1 if its argument is true and zero otherwise. Similarly, Zhu *et al.* (2010) provided Cumulative Mean Estimation (CUME) which is variation of SIR and based on  $E[E\{\mathbf{z}I(Y \leq \tilde{Y})\}E\{\mathbf{z}I(Y \leq \tilde{Y})\}^\top]$ . Neither method requires choosing  $H$  and Shaker & Prendergast (2011) showed that they can be successfully combined to obtain excellent results. Consequently, we will also consider CUME following transformation based on the first CUVE direction.

For SIR and SAVE the user must specify the subranges of  $Y$  used for ‘slicing’ with the easiest approach to set  $H$  equally probable slices. In practice this equates to ordering the data by the magnitude of the response and allocating an (approximately) equal number of observations to each slice. For estimation we use Rs `dr` package (Weisberg, 2002) which, by default, uses  $\max(8, p + 3)$  for  $H$ .

An advantage of SIR, SAVE, CUME and CUVE is that the  $y_i$ 's are used only to allocate  $\mathbf{x}_i$ 's. Consequently, we would not expect extremely large  $y_i$ 's to have the same detrimental effect on estimation as they do for OLS. To highlight this we reconsider Model 3 and adopt Method 2 as follows. SAVE is used to get the initial estimate to the direction of  $\boldsymbol{\beta}$ . SIR

Method	$n = 100$	$n = 200$	Method	$n = 100$	$n = 200$
SIR	0.066	0.065	SAVE SIR (M2)	0.247	0.835
	<i>0.106</i>	<i>0.107</i>		<i>0.292</i>	<i>0.266</i>
CUME	0.070	0.068	SAVE SIR (M2-it)	0.363	0.934
	<i>0.090</i>	<i>0.090</i>		<i>0.401</i>	<i>0.234</i>
SAVE	0.174	0.585	CUVE CUME (M2)	0.972	0.993
	<i>0.190</i>	<i>0.237</i>		<i>0.034</i>	<i>0.003</i>
CUVE	0.889	0.957	CUVE CUME (M2-it)	0.987	0.995
	<i>0.053</i>	<i>0.016</i>		<i>0.021</i>	<i>0.002</i>

Table 3: Average  $\text{cor}^2(\mathbf{X}\boldsymbol{\beta}, \mathbf{X}\hat{\mathbf{b}})$  across 10,000 simulated data sets for Model 3 with two different choices of  $n$  and  $p = 20$  and where  $\hat{\mathbf{b}}$  is the estimate from various methods. M2 refers to transformation Method 2 and M2-it refers to this transformation with iterative estimation. Standard deviations are in italics.

is then used on either the transformed  $y_i$ 's (Method 1) or  $\mathbf{x}_i$ 's (Method 2) where the SAVE direction has been used to facilitate the transformations. Similarly, we use CUME following a transformation using the CUVE estimated direction. We also consider the iterative estimation procedures for both. In Table 3 we provide the results from 10,000 simulations where  $p = 20$  and  $n = 100$  or 200. Due to symmetry, both SIR and CUME fail to estimate the direction of  $\boldsymbol{\beta}$ . SAVE also has trouble estimating the direction of  $\boldsymbol{\beta}$ , especially for  $n = 100$ , although improvements are found for  $n = 200$  and we observed good results for  $n = 500$  (not shown). CUVE, on the other hand, performs well, even for  $n = 100$ . The results also indicate that transformation Method 2 results in improved estimation although the combination of SAVE and SIR is only successful for the larger sample size. The combination of CUVE and CUME, however, provides excellent results even for  $n = 100$ . For both approaches, the iterative estimation scheme also provides improvements, as evidenced by the increase mean squared correlations.

## 5.2 Detecting multiple directions with OLS

Garnham & Prendergast (2013b) showed that OLS can be used to find multiple directions when different response transformations are employed. They obtain several OLS slopes with weights related to a leave-one-out sensitivity and then obtain one or more directional estimates for the  $\boldsymbol{\beta}_k$ 's. Similarly, we show here that a simple two-step estimation procedure for OLS can work exceptionally well when OLS is faced with the task of finding two directions, one of which is expected to be non-detectable due to symmetric dependency. As a matter of

comparison we will also consider PHD|OLS which is an iterative version of PHD and OLS considered by Shaker & Prendergast (2011). Here, the first direction estimated is the OLS slope. Then PHD is used conditional on this OLS slope estimate already been detected so that only new information is found in the second direction. The model we will focus on is given below which was also considered by Shaker & Prendergast (2011).

**Model 4.**  $Y = \sin(0.5\beta_1^\top \mathbf{x}) + \cos(0.5\beta_2^\top \mathbf{x}) + 0.3\varepsilon$  where  $\beta_1 = [1, 2, -3, 0, \dots, 0]^\top$  and  $\beta_2 = [1, 1, 0, -2, 0, \dots, 0]^\top$ .

For the model above we will consider the performance of SIR, PHD, PHD|OLS and three new approaches based on Methods 1 and 2. For these new approaches we will use the OLS slope vector as the first estimated direction and then use the transformation methods to estimate a second direction.

Method	$n = 100$		$n = 200$		$n = 500$		$n = 1000$	
	$\bar{r}_1$	$\bar{r}_2$	$\bar{r}_1$	$\bar{r}_2$	$\bar{r}_1$	$\bar{r}_2$	$\bar{r}_1$	$\bar{r}_2$
SIR	0.776	0.277	0.872	0.326	0.952	0.448	0.977	0.615
	<i>0.142</i>	<i>0.199</i>	<i>0.088</i>	<i>0.230</i>	<i>0.028</i>	<i>0.269</i>	<i>0.013</i>	<i>0.264</i>
PHD	0.928	0.410	0.967	0.422	0.987	0.428	0.994	0.431
	<i>0.046</i>	<i>0.231</i>	<i>0.020</i>	<i>0.235</i>	<i>0.007</i>	<i>0.237</i>	<i>0.003</i>	<i>0.236</i>
PHD OLS	0.929	0.678	0.967	0.832	0.988	0.931	0.994	0.965
	<i>0.048</i>	<i>0.207</i>	<i>0.020</i>	<i>0.109</i>	<i>0.007</i>	<i>0.038</i>	<i>0.003</i>	<i>0.018</i>
OLS,M1	0.942	0.716	0.976	0.850	0.991	0.934	0.996	0.966
	<i>0.045</i>	<i>0.172</i>	<i>0.017</i>	<i>0.086</i>	<i>0.006</i>	<i>0.037</i>	<i>0.003</i>	<i>0.018</i>
OLS,M2	0.949	0.722	0.980	0.852	0.993	0.935	0.996	0.967
	<i>0.038</i>	<i>0.171</i>	<i>0.013</i>	<i>0.085</i>	<i>0.004</i>	<i>0.036</i>	<i>0.002</i>	<i>0.017</i>

Table 4: Average first and second canonical correlations ( $\bar{r}_1, \bar{r}_2$ ) between  $\mathbf{X}[\beta_1, \beta_2]$  and  $\mathbf{X}\hat{\mathbf{B}}$  across 10,000 simulated runs for data generated from Model 4 with different choices of  $n$  and  $p$  and where  $\hat{\mathbf{b}}$  is the estimate from one of five methods; OLS, PHD, Method 1 (M1) and Method 2 (M2). Standard deviations are in parentheses.

In Table 4 we provide the simulated average first and second canonical correlations between  $\mathbf{X}[\beta_1, \beta_2]$  and  $\mathbf{X}\hat{\mathbf{B}}$  where  $\hat{\mathbf{B}}$  is a  $p \times 2$  matrix consisting of the first and second estimated directions. A large average first canonical correlation,  $\bar{r}_1$ , indicates that the approach successfully detects the first direction. Similarly, a large  $\bar{r}_2$  is indicative of good performance in detecting the second direction. SIR and PHD each are capable of finding one of the directions - for SIR the direction it is expected to find  $\beta_1$  and for PHD it is  $\beta_2$ . However, these methods do not perform well at finding the other direction. PHD|OLS is expected to find both

and the average canonical correlations indicate this, although the method may have some trouble in estimating the second direction for  $n = 100$ . The transformation methods provide improvements in estimating both directions; certainly with respect to SIR and PHD and marginally better results than even PHD|OLS.

We could similarly use robust  $M$ -estimation regression methods here too. Rather than repeat the simulation for similar results, we will choose this approach for the example considered in the next section.

## 6 The Ozone data example

Li (1992) considered the Ozone data from Breiman & Friedman (1985) which consists of 330 observations and eight predictors (e.g. wind speed, humidity etc., refer to Table 4 of Li 1992 for full list of predictors). The response is atmospheric ozone concentration. Li (1992) notes that the method Sliced Inverse Regression SIR finds a quadratic relationship (although not one that includes symmetric dependency) between the response and eight predictors and that almost an identical relationship can be found using least squares. Using PHD, another direction is found that eluded SIR and which provides an ESSP that exhibits symmetric dependency.

In this example we will also consider the Ozone data example. Let  $Y$  denote the response variable and  $\mathbf{x}$  denote the eight-dimensional vector of predictor variables. As previously we let the sample data be denoted  $\{y_i, \mathbf{x}_i\}_{i=1}^{330}$ . We will base our model on  $\sqrt{Y}$  which, as we will see shortly, allows methods such as OLS,  $M$ -estimator regression methods and SIR to detect a linear relationship between the response and predictors. We will also use robust  $M$ -estimator as a robust least-squares method with the Huber weight function in the analysis. For convenience we will refer to this method as RR.

Let  $\hat{\mathbf{b}}_1$  be the estimated slope for the RR regression of the  $\sqrt{y_i}$ 's on the  $\mathbf{x}_i$ 's. A plot of the  $\sqrt{y_i}$ 's versus the  $\hat{\mathbf{b}}_1^\top \mathbf{x}_i$ 's in Plot A of Figure 3 shows a linear relationship between the response and the dimension reduced predictors (labelled '1st RR dr predictor' on the plot). We now use transformation Method 1 with the first PHD direction but with RR replacing OLS and let  $\hat{\mathbf{b}}_2$  denote this new estimate. Plot B shows that RR has now found another direction exhibiting symmetric dependency. We now use OLS to fit a model to the  $\hat{\mathbf{b}}_1^\top \mathbf{x}_i$ 's, the  $\hat{\mathbf{b}}_2^\top \mathbf{x}_i$ 's and the square of each of these (we did not include the multiple of the two for a full quadratic model since this had little contribution). The estimated model is

$$\widehat{Y^{1/2}} = -212.89 + 1.22 \times (\hat{\mathbf{b}}_1^\top \mathbf{x}) + 13.94 \times (\hat{\mathbf{b}}_2^\top \mathbf{x}) + 0.18 \times (\hat{\mathbf{b}}_1^\top \mathbf{x})^2 - 0.22 \times (\hat{\mathbf{b}}_2^\top \mathbf{x})^2.$$

The above fitted model explains approximately 75% of the variation in square-root of the

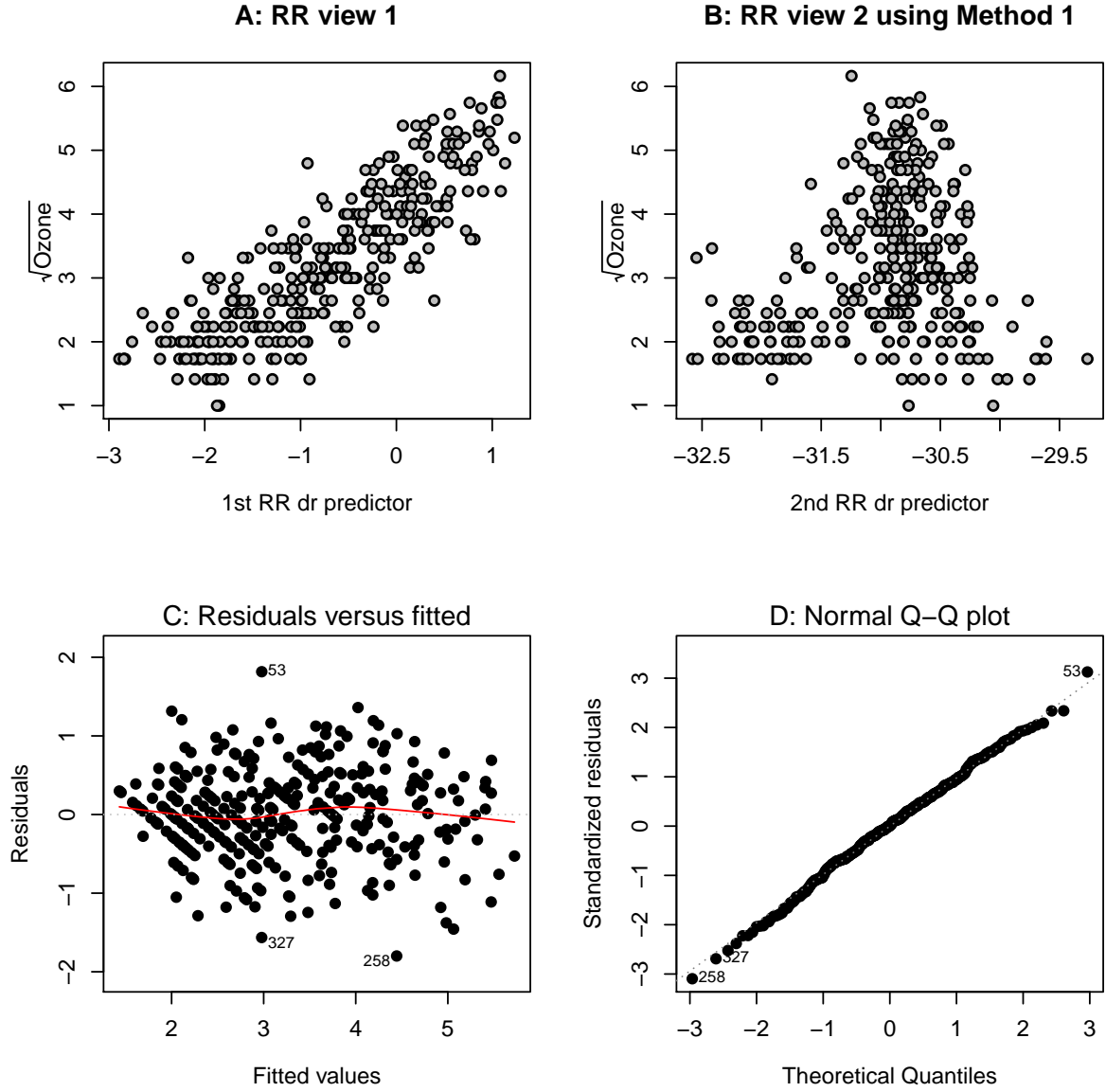


Figure 3: Plots of (A) the ESSP found by RR, (B) an ESSP created using a second direction found using RR following transformation Method 1, (C) residuals versus fits of a least squares fit to the dimension reduced predictors from Plots (A) and (B) and the square of these dimension reduced predictors and (D) the corresponding normal Quantile-Quantile plot.

response indicating a good fit and all of the terms in the model were highly significant. In Plots C and D we provide the residuals versus fits plot for the fit and also the Quantile-Quantile plot to check to see whether one could assume something close to a normal error term for the underlying model. These plots are excellent indicating that if we were to assume



a normally distributed error term with homogeneous variance then there is no evidence here to suggest that such an assumption would not hold approximately. Consequently, we have successfully used RR twice to find two directions that can be used to construct a simple model with simple error term properties.

## 7 Discussion

This paper showed that simple response and predictor transformations can be used to remove the problem of symmetric dependency that effects some dimension reduction methods. While we initially showed that OLS and PHD can be successfully employed in tandem for improved estimates, our approaches need not be limited to these methods. To highlight this we also showed the popular robust  $M$ -estimation methods can be used as well as Sliced Inverse Regression in conjunction with Sliced Average Variance Estimates and associated cumulative slicing approaches. These approaches are particularly useful when faced with very large response values that can be detrimental to OLS estimation. Another interesting outcome from this paper was the ability in which OLS, and robust equivalents, could be used to find more than one direction.

## A Proof of Theorem 1

Throughout let  $E(\mathbf{x}) = \boldsymbol{\mu}$  and  $\text{Var}(\mathbf{x}) = \boldsymbol{\Sigma}$  and recall  $\mathbf{v} = c\boldsymbol{\beta}$ . It can be shown that (see, for e.g., Prendergast, 2005) Condition 1 is equivalent to

$$E(\mathbf{x}|\mathbf{v}^\top \mathbf{x}) = \boldsymbol{\mu} + \{(\mathbf{v}^\top \boldsymbol{\Sigma} \mathbf{v})^{-1} \mathbf{v}^\top [E(\mathbf{x}|\mathbf{v}^\top \mathbf{x}) - \boldsymbol{\mu}]\} \boldsymbol{\Sigma} \mathbf{v} \quad (11)$$

Since  $E[\text{sign}(\mathbf{v}^\top \mathbf{x} - \mathbf{v}^\top \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})] = E\{\text{sign}(\mathbf{v}^\top \mathbf{x} - \mathbf{v}^\top \boldsymbol{\mu})E(\mathbf{x} - \boldsymbol{\mu}|\mathbf{v}^\top \mathbf{x})\}$ , then from (11),

$$E[t_x(\mathbf{x} - \boldsymbol{\mu}; \mathbf{v})] = E[\text{sign}(\mathbf{v}^\top \mathbf{x} - \mathbf{v}^\top \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})] = \{(\mathbf{v}^\top \boldsymbol{\Sigma} \mathbf{v})^{-1} \mathbf{v}^\top E[t_x(\mathbf{x} - \boldsymbol{\mu}; \mathbf{v})]\} \boldsymbol{\Sigma} \mathbf{v} = c_4 \boldsymbol{\Sigma} \mathbf{v} \quad (12)$$

where we identify here that  $c_4 \in \mathbb{R}$ .

Similarly, by noting  $E[\text{sign}(\mathbf{v}^\top \mathbf{x} - \mathbf{v}^\top \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})Y] = E\{\text{sign}(\mathbf{v}^\top \mathbf{x} - \mathbf{v}^\top \boldsymbol{\mu})YE(\mathbf{x} - \boldsymbol{\mu}|\mathbf{v}^\top \mathbf{x})\}$  since, from the model in (1),  $Y$  is a function of  $\boldsymbol{\beta}^\top \mathbf{x}$  and  $\varepsilon$  where  $\varepsilon$  is independent of  $\mathbf{x}$ , we can also show that

$$\text{Cov}[\text{sign}(\mathbf{v}^\top \mathbf{x} - \mathbf{v}^\top \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu}), Y] = c_5 \boldsymbol{\Sigma} \mathbf{v} \quad (13)$$

for a  $c_5 \in \mathbb{R}$ . This shows that (6) holds.

Now, using (12),

$$\text{Var}[t_x(\mathbf{x} - \boldsymbol{\mu}; \mathbf{v})] = \boldsymbol{\Sigma} - c_4^2 \boldsymbol{\Sigma} \mathbf{v} \mathbf{v}^\top \boldsymbol{\Sigma}$$

since  $E[t_x(\mathbf{x} - \boldsymbol{\mu}; \mathbf{v})t_x(\mathbf{x} - \boldsymbol{\mu}; \mathbf{v})^\top] = E[(\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^\top] = \boldsymbol{\Sigma}$ . Therefore (for e.g., use the Small Rank Adjustment Lemma, Horn & Johnson, 1985, page 19)

$$\text{Var}[t_x(\mathbf{x} - \boldsymbol{\mu}; \mathbf{v})] = \boldsymbol{\Sigma}^{-1} + \frac{c_4^2}{1 - c_4^2 \mathbf{v}^\top \boldsymbol{\Sigma} \mathbf{v}} \mathbf{v} \mathbf{v}^\top$$

In conjunction with (13), this shows that (7) also holds completing the proof.

## References

- BREIMAN, L., & FRIEDMAN, J. H. 1985. Estimating optimal transformations for multiple regression and correlation. *J. Amer. Statist. Assoc.*, **80**, 580–619. With discussion and with a reply by the authors.
- BRILLINGER, D. R. 1977. The identification of a particular nonlinear time series system. *Biometrika*, **64**, 509–515.
- BRILLINGER, D. R. 1983. A generalized linear model with “Gaussian” regressor variables. *Pages 97–114 of: A Festschrift for Erich L. Lehmann*. Wadsworth Statist./Probab. Ser. Belmont, CA: Wadsworth.
- COOK, R. D. 1998a. *Regression graphics*. New York: John Wiley & Sons Inc. Ideas for studying regressions through graphics.
- COOK, R. D., & WEISBERG, S. 1991. Discussion of “Sliced inverse regression for dimension reduction.”. *J. Amer. Statist. Assoc.*, **86**, 328–332.
- COOK, R. DENNIS. 1998b. Principal Hessian directions revisited. *J. Amer. Statist. Assoc.*, **93**, 84–100. With comments by Ker-Chau Li and a rejoinder by the author.
- FOX, J., & WEISBERG, S. 2011. *An R Companion to Applied Regression*. SAGE Publications.
- GARNHAM, A. L., & PRENDERGAST, L. A. 2013a. A note on least squares sensitivity in single-index model estimation and the benefits of response transformations. *Electron. J. Stat.*, **7**, 1983–2004.
- GARNHAM, A. L., & PRENDERGAST, L. A. 2013b. Optimal response transformations for single- and multi- index models. *Submitted*.
- HALL, P., & LI, K.-C. 1993. On almost linearity of low-dimensional projections from high-dimensional data. *Ann. Statist.*, **21**, 867–889.

- HORN, R. A., & JOHNSON, C. A. 1985. *Matrix analysis*. New York: Cambridge University Press.
- HUBER, P. J. 1964. Robust estimation of a location parameter. *Ann. Math. Stat.*, **35**, 73–101.
- HUBER, P. J. 1973. Robust regression: asymptotics, conjectures and Monte Carlo. *Ann. Statist.*, **1**, 799–821.
- LEEB, H. 2013. On the conditional distributions of low-dimensional projections from high-dimensional data. *Ann. Stat.*, **41**(2), 464–483.
- LI, K.-C. 1991. Sliced inverse regression for dimension reduction. *J. Amer. Statist. Assoc.*, **86**, 316–342. With discussion and a rejoinder by the author.
- LI, K.-C. 1992. On principal Hessian directions for data visualization and dimension reduction: another application of Stein’s lemma. *J. Amer. Statist. Assoc.*, **87**, 1025–1039.
- LI, K.-C., & DUAN, N. 1989. Regression analysis under link violation. *Ann. Statist.*, **17**, 1009–1052.
- LUE, H.-H. 2001. A study of sensitivity analysis on the method of principal hessian directions. *Computation. Stat.*, **16**(1), 109–130.
- PRENDERGAST, L. A. 2005. Influence functions for sliced inverse regression. *Scand. J. Statist.*, **32**, 385–404.
- PRENDERGAST, L. A. 2008. Trimming influential observations for improved single-index model estimated sufficient summary plots. *Comput. Statist. Data Anal.*, **52**, 5319–5327.
- PRENDERGAST, L. A., & SHEATHER, S.J. 2013. On sensitivity of inverse response plot estimation and the benefits of a robust estimation approach. *Scand. J. Stat.*, **40**, 219–237.
- PRENDERGAST, L. A., & SMITH, J. A. 2010. Influence functions for dimension reduction methods: an example influence study of principal Hessian direction analysis. *Scand. J. Stat.*, **37**, 588–611.
- R CORE TEAM. 2013. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.
- SHAKER, A. J., & PRENDERGAST, L. A. 2011. Iterative application of dimension reduction methods. *Electron. J. Stat.*, **5**, 1471–1494.

- VENABLES, W. N., & RIPLEY, B. D. 2002. *Modern Applied Statistics with S*. Fourth edn. New York: Springer. ISBN 0-387-95457-0.
- WEISBERG, S. 2002. Dimension reduction regression in R. *J. Stat. Softw.*, **7**(1), 1–22.
- ZHU, LI-PING, ZHU, LI-XING, & FENG, ZHENG-HUI. 2010. Dimension reduction in regressions through cumulative slicing estimation. *J. Amer. Statist. Assoc.*, **105**(492), 1455–1466.